Wilson KJ.

An investigation of dependence in expert judgement studies with multiple experts.

International Journal of Forecasting **(2016)**

DOI: http://dx.doi.org/10.1016/j.ijforecast.2015.11.014

An investigation of dependence in expert judgement studies with multiple experts

Abstract

Expert judgement plays an important role in forecasting and elsewhere as it can be used to quantify models when no data are available and to improve predictions from models when combined with data. In order to provide defensible estimates of unknowns in an analysis the judgements of multiple experts can be elicited. Mathematical aggregation methods can be used to combine these individual judgements into a single judgement for the decision maker. However, most mathematical aggregation methods assume judgements coming from experts that are independent. This is unlikely to be the case in practice. This paper investigates dependence in expert judgement studies, both within and between experts. It gives the most comprehensive analysis to date by considering all studies in the TU Delft database. It then assesses the practical significance of the dependencies identified in the studies by comparing the performance of several mathematical aggregation methods with varying dependence assumptions. Between expert correlations were more prevalent than within expert correlations. For studies which contained between expert correlations, models which include these improved forecasts. The implications for the use of expert judgement in forecasting are discussed.

Keywords: Expert judgement, dependency, mathematical aggregation, Bayesian methods, finance

1 Introduction

Expert judgement has been used in forecasting informally when data aren't available and formally to bound problems, qualitatively structure models and quantify unknowns within models. One example in DeWispelare et al [14] used expert judgements to estimate the probabilities within high-level nuclear waste regulation. Probabilistic forecasts were elicited from five climatologists on parameters of complex climatic models. In a very different application,

Brandt et al [4] considered the evaluation of forecasts in political conflict dynamics. They considered forecasts from analysts of densities and the importance of taking uncertainty in forecasts into account when evaluating models. Jochmann et al [19] considered the use of expert judgement VAR forecasting in combination with data. They identified model parameters over which prior uncertainty distributions could be elicited from experts and indicated that this approach offers improvements in accuracy over a solely data driven model. As we see, expert judgement can be used on its own when there are no data available or in combination with data in a Bayesian analysis. In both cases we require a procedure to obtain the expert judgements we require.

An important question in any expert judgement study is whose judgements to elicit. That is, what constitutes an expert? There are various views on this. In a pure subjectivist Bayesian analysis then the expert could simply be the person from whom the unknowns are being elicited. However, if we are considering the expert problem [15] in which experts are being asked for advice from a specific decision maker, then the role of expert will require more justification. Instead, we may use a definition from [16] that experts are "persons to whom society and/or his peers attribute special knowledge about the matters being elicited". Crucially, it is also the ability to use this knowledge that defines a good expert [28]. For discussion on the selection of experts see Sections 1.3 and 5 of [16]. In the expert problem multiple experts are typically used to improve the information given to the decision maker.

When performing elicitation, questions must always be asked about quantities which are relevant to the expert, rather than abstract model parameters. Typically questions are put to experts about probabilities or quantiles and then these are converted to the parameters of probability distributions by the analyst. However, there is much evidence in the psychological literature that humans are susceptible to heuristics and biases when giving quantitative assessments such as these and efforts must be made in any elicitation to minimise the influence of these biases. Three common heuristics which can lead to such biases are judgement by representativeness; evaluating probabilities of events based on how similar two things are while ignoring base rates, judgement by availability; basing probabilities of events on how easily the events can be recalled and anchoring; the expert is given an irrelevant value for the probability of an event and then inadequately adjusts this up or down based on how likely they think this event is. In particular, availability and representativeness can shift elicited probabilities and anchoring can result in quantiles which are too narrow, resulting in incorrect probability distributions. For more information on heuristics and biases see [21, 31].

When multiple experts give judgements in a study it may be necessary, or at least preferable, to combine their judgements into a single coherent judgement

to report back to the decision maker. There are two main ways to do this: behavioural aggregation and mathematical aggregation. In behavioural aggregation the experts are typically brought together into a single place and the objective is to come to a consensus about each quantity in the analysis. When is is not possible to bring experts together, and to avoid biases resulting from freely interacting groups, methods such as the Delphi technique have been developed which involve interaction between experts under the control of the analyst [29]. For further information on behavioural aggregation see [13, 22] and for a recent discussion of behavioural versus mathematical approaches see [5]. In mathematical aggregation a mathematical rule is used to combine the judgements of the experts. There are two main ways of doing this: opinion pools and Bayesian aggregation. In opinion pools weights are given to each expert and then judgements are combined linearly or logarithmically using these weights. The weights can be specified based on performance of the experts on questions to which the analyst knows the answer, the judgements of the decision maker, self weighting by the experts or equal weights. In Bayesian aggregation the expert judgements are regarded as data and are combined using Bayes theorem.

The majority of the mathematical aggregation methods proposed in the literature assume that the judgements of experts are independent, both of other judgements made by that expert, and of judgements made by other experts [15]. In practice this may not be the case, as individual experts may be subject to the same biases consistently, different experts may be subject to the same biases and different experts may have similar backgrounds and experience. Thus it seems likely that dependencies will exist within expert judgement studies and therefore the impact of these on model accuracy should be assessed.

There are some models in the literature which consider correlations, or whose modelling could include correlations, with multiple experts. Most are in the Bayesian aggregation literature and require the decision maker to specify the dependencies between experts and biases of experts. Examples include [23, 34, 20]. More examples are given in [15]. An alternative to the decision maker specifying these values is to offer the decision maker the empirical values of the correlations from seed questions which the analyst knows the answer to but the experts do not.

In this paper we consider the data from 45 expert studies in which the judgements of multiple experts in various fields from the nuclear sector to health and banking were elicited. The studies were conducted by TU Delft and released as part of [10]. A fuller description is given in Section 4.2. For each study in the data set we investigate whether within expert and between expert correlations are present for all of the seed variables in that study. This provides the most comprehensive analysis to date of the extent and type of

dependence within expert judgement studies. Further, we also fit several of the most commonly used mathematical aggregation methods from the literature to each of these studies and evaluate their accuracy using a number of metrics. This allows us to make some general comments both about the types of correlations which are present in expert judgement studies and also about whether these correlations are having a practically significant effect on the accuracy of the predictions resulting from the models.

The rest of the paper is structured as follows. In Section 2 we review the two main approaches to mathematical aggregation, Bayesian aggregation and opinion pooling. In Section 4.1 we consider the different possible sources of correlation within expert judgement studies and how they might be measured and in Section 4.2 we provide details of the expert judgement database from TU Delft. In Section 5 we evaluate the dependence present in the TU Delft studies and in Section 6 we fit a number of mathematical aggregation models to the case studies and evaluate the accuracy of each model for each study. We conclude the paper in Section 7 with a summary and discussion.

2 Mathematical Expert Judgement Approaches

2.1 Bayesian Aggregation

Suppose we are interested in an event or unknown quantity which we shall call θ. Then the experts will give us, through an elicited probability or quantiles of θ, individual prior distributions, $f_{0,i}(\theta)$, for experts $i = 1, \ldots, E$. The set of these elicited distributions is $\underline{D} = (f_{0,1}(\theta), \ldots, f_{0,E}(\theta))$. A Bayesian aggregation method then works by applying Bayes Theorem,

$$f_{1,DM}(\theta \mid \underline{D}) \propto f_{0,DM}(\theta) L_{DM}(\underline{D} \mid \theta),$$

where $f_{0,DM}(\theta)$ represents the decision maker's prior probability distribution for unknown θ, $L_{DM}(\underline{D} \mid \theta)$ is the decision maker's likelihood of observing $\underline{D}$ given θ and $f_{1,DM}(\theta \mid \underline{D})$ is the decision maker's posterior distribution for θ.

The main challenge in this method is eliciting from the decision maker the likelihood function $L_{DM}(\underline{D} \mid \theta)$. It is in this likelihood function that the correlations in the expert judgement study can be captured. We see that the outcome of Bayesian aggregation methods is a subjective probability distribution which gives the updated beliefs of the decision maker in the tradition of a subjectivist Bayesian analysis.

Some common Bayesian aggregation methods, which will be used for the comparison study later in the paper, are an approach based on the multivariate

Normal distribution proposed by Winkler in [34] and a copula approach proposed by Jouini and Clemen [20]. Please refer to these references for further information on the methods.

2.2 Opinion Pooling

Opinion pooling aims to give weights to individual experts. The decision maker's distribution for the unknown quantity θ is then the weighted average of all of the experts' judgements for that quantity.

Suppose that expert i gives elicited values which result in the distribution $f_i(\theta)$ and that the weight attached to expert i is w_i, $0 \leq w_i \leq 1$, where $\sum_{i=1}^{E} w_i = 1$. Define $\underline{D} = (f_1(\theta), \ldots, f_E(\theta))$ to be the set of expert distributions for θ. The decision maker's consensus distribution will take one of two forms, a linear pool

$$f(\theta) = \sum_{i=1}^{E} w_i f_i(\theta),$$

or a logarithmic pool,

$$f(\theta) = k \prod_{i=1}^{E} f_i(\theta)^{w_i},$$

where k is a normalising constant to ensure that the distribution integrates to 1. Note that in the logarithmic pool if any expert gives θ a probability of 0 then it will have a probability of 0 in the consensus distribution.

Common opinion pooling methods to be used in the comparison are the Classical Method of Roger Cooke and others [8] and an approach proposed by Babuscia and Cheung [2].

3 Combining Model Forecasts

The work in this paper is related to the literature on combining model forecasts. For good overviews of the topic see [7, 1]. A great deal of work has been carried out which indicates that combining model forecasts improves forecast accuracy in comparison to single forecasts. Examples include [24, 32, 18]. In particular, research has shown that we can achieve dramatic forecasting improvements by averaging forecasts, whether those forecasts be outputs of data driven models or achieved by some other means.

Work in this area has looked at ideas such as minimum variance models, Bayesian combinations and ARIMA models [7]. There is a body of evidence from this context which suggests that using comparatively simple averaging

methods which ignore correlations in their estimation can improve the accuracy of forecasts [27, 32]. Similarly, there is evidence that simple averaging (equal weights) often outperforms sophisticated models [18].

A subset of the combining model forecasts literature has considered the combination of judgemental forecasts from, typically, experts. There has been work carried out on achieving consensus amongst judgements, simple averaging of judgemental forecasting, bootstrapping to model judgements using data and work looking at the optimal number of experts to consult when combining judgemental forecasts [7, 26]. Similarly, there has been work on the combination of forecasts using the Bayesian approach, considering the combination of the forecasts as the posterior distribution in a Bayesian analysis [25]. A question which is still not resolved in the literature is that of whether it is optimal to use hard or soft methods to combine forecasts.

We can use the discussion above to place the current study within the forecasting literature. We consider correlations in expert judgements. We identify, from the largest known database of expert judgement studies, which types of correlation are present. We then use a number of well known combination methods for each study and compare the predictions resulting from the methods first to equal weights and then to eachother for all studies and then just the studies in which there are high correlations. Therefore we are contributing to the literature (i) an assessment of the extent of correlations within a large number of real judgemental studies, (ii) whether more sophisticated methods for combining judgements outperform simple averaging (or not) and (iii) whether methods taking into account correlations outperform those which don't (or not).

4 Correlations in Group Expert Judgement Studies

4.1 Sources of Correlation

There are several different areas within mathematical group expert judgement models in which there are potential correlations. Such correlations have the potential to affect the accuracy of opinion pooling methods and, if each type of correlation exists in a study, should be captured in Bayesian aggregation methods.

In order to identify the potential correlations present in expert judgement studies, it will be useful to consider two types of uncertainty which are relevant to experts making judgements. The first is *aleatory uncertainty*, which represents randomness in the state of the world. For example, if we were to roll

a die, we are uncertain as to how many spots will end face up. It doesn't matter how many times we have rolled the die in the past, this will always be an uncertain event. The second type we shall consider is *epistemic uncertainty*, which represents our own lack of knowledge. For example, if someone were to hand us a loaded die then we would have additional uncertainty around how likely we are to see a six, for example. We could reduce our uncertainty about this event by rolling the die a large number of times and counting the number of sixes.

The possible correlations within an expert judgement study are:

1. Correlation between the experts for individual quantities: these could be a result of the similar past experience and common knowledge of the experts or because the experts are susceptible to the same biases through their use of heuristics.

2. Correlation within individual experts' assessments of different quantities: these could be as a result of a consistent susceptibility of an expert to the same biases.

3. Correlation between the experts for different quantities: these could be as a result of multiple experts being consistently susceptible to the same biases.

4. Underlying aleatory correlation between the values of the quantities to be assessed in the expert judgement study: plotting one against another there is a relationship.

The first three types of correlation are conditional on the true value of the underlying variable. In the Bayesian mathematical aggregation methods there are a further two possible correlation types. They are:

5. Underlying epistemic correlation between the quantities in the study: learning about one quantity will inform us as to the likely value of another.

6. Correlation between the experts' judgements and the decision maker's judgements [15]. These could again come about as a result of common knowledge or susceptibility to the same biases.

It seems plausible that a combination of some or all of these correlations are present in all expert judgement studies. The methods which can be used to assess them differ depending on the type of correlation in question, however.

The correlations between experts for individual quantities and within individual experts for multiple quantities can be assessed empirically for a given study using seed variables, questions which are related to the current problem but for which the answers are known to the analyst, as long as experts are

being asked to give values for multiple quantities. This can then be built into aggregation methods.

The underlying correlations as a result of aleatory uncertainty between the values of the assessed quantities can also be assessed empirically if the quantities are to be observed multiple times (i.e. could be plotted on a scatter graph), for example as multiple identical components on test. The epistemic correlations on the values of the quantities can be elicited from individual experts by asking questions about their uncertainty over multiple unknowns simultaneously.

Simply because these correlations exists within a study it does not necessarily mean that they are having an influence on the accuracy of the outputs of an aggregation method, however. There is of course a difference between statistical significance and practical importance.

Consider the methods identified above. The Classical method and Babuscia and Cheung method make the assumption that different experts are giving independent answers for questions. That is, they assume that correlations of types 1 and 3 above are not present. The use of seed questions makes the implicit assumption that an individual expert's performance on seed questions would be similar to their performance on questions of interest. Therefore they are all assuming a strong correlation of type 2. They are unaffected by correlation types 4,5 and 6 as repeated measurements are not considered and the decision maker is not involved in the aggregation process.

The multivariate Normal and Copula models include the correlations between experts for the same question (type 1) explicitly. Similarly, both models allow the decision maker to define a "bias" for an individual expert's distribution which takes into account correlation type 2. As the models discusses experts eliciting uncertainty for a single quantity, they do not include correlations of types 3, 4 or of 5. They do, however, include correlations of type 6.

4.2 TU Delft Expert Judgement Database

In [10], the authors explored the database of all expert judgement studies conducted by TU Delft. In all the studies involved over 67,000 expert probability distributions. The studies have been conducted over several different sectors, the most prominent being the nuclear industry but other notable sectors where a number of studies have been carried out are the chemical and gas industry, the aerospace sector, health, banking, volcanoes and dams.

In each study a number of seed variables are elicited from each expert. The experts are usually asked for their median and 5% and 95% quantiles for each variable, though in a few cases other quantiles are asked for in addition. The experts are then asked for the same quantiles for the variables of interest for

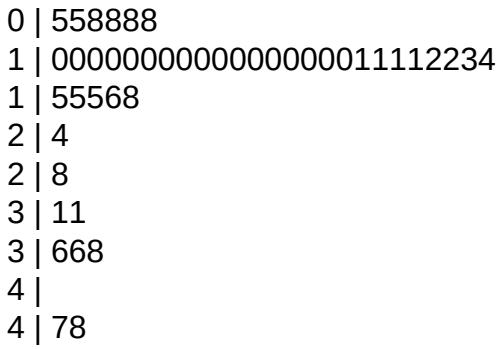

Figure 1: Stem and leaf plot showing the number of seed questions in each of the Delft expert judgement studies.

which the answers are unknown and, usually, will not become known in the future.

To illustrate the form that the data take, we will consider one specific study from the database. The study concerned the failure frequency of gas pipelines in the Netherlands and was conducted by TU Delft with Gasunie, the national gas company [11]. Gasunie was interested in this in order to make maintenance and safety decisions about their pipeline network. Fifteen experts took part in the study, from Gasunie and a number of similar organisations in other countries. In order to make decisions, Gasunie required one single set of probability distributions on future failure frequencies of gas pipelines. Each expert was asked questions on seventeen seed variables, which were taken from overall population data on third party interference and damages found on a large study of Dutch pipelines. It is these seed variables which we will use in the analyses in this paper. This means that from this study we have 17x15 medians, 5% and 95% quantiles from experts for questions to which we know the true value. Thus we are able to assess how well each expert answered each question.

We will analyse the data from 45 studies from all experts in order to assess the dependence present in the TU Delft expert judgement database. A summary of the numbers of seed questions and experts are given in Figures 1 and 2. We see from the plots that there are typically between 5 and 18 seed questions asked to each expert in a study with a maximum of 48. On average, the number of experts used in a study is smaller than the number of seed questions used, often less than 10, and the maximum number of experts used is 77.

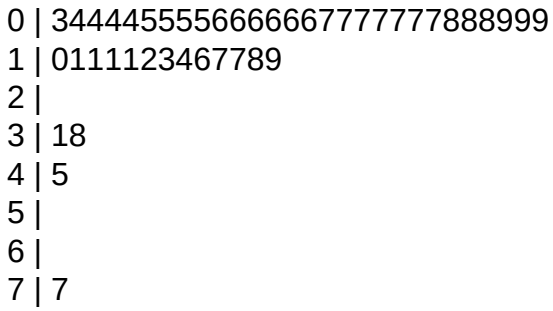

Figure 2: Stem and leaf plot showing the number of experts in each of the Delft expert judgement studies.

4.3 The Current Analysis

In this paper we are concerned with an analysis of 45 expert judgement studies conducted in Delft University of Technology and released as part of [10]. Within each these studies we will empirically estimate

- The correlations between the different experts for the seed questions in each study.
- The correlations between the different assessments of individual experts within each study.

The other correlations are outside the scope of this analysis.

We can also assess the effect that correlations are having on the accuracy of different mathematical aggregation models. We fit the aggregation models identified earlier to the data from each of the studies and assess the accuracy of the combined expert or decision maker for each method. Aggregation methods which assume various forms of independence and dependence are investigated to help to give an assessment of the practical significance of the various correlations which are found to be present.

There is a question when analysing data consisting of seed variables as to how consistently expert performance (and biases, correlations, etc) for seed variables corresponds to performance for the real questions in expert judgement studies. We simply comment here that TU Delft work hard to ensure that seed questions are as closely related to the variables of interest as possible and are in the relevant field of the study. For further discussion on the studies in question see [10], and for a general discussion around seed variables in expert judgement studies see [5], accompanying commentaries [9, 35, 33] and response [6].

5 Dependence in TU Delft Studies

5.1 Analysis of Studies

We initially consider as a whole every assessment from every expert in 45 studies. We assume that the median given by an expert for a seed variable is that expert's "best guess" at the true value of the variable. Doing so, we can then plot each of these best guesses from each expert over the 45 studies against the true value of the seed variable in each case. In practice we use the natural logarithmic scale. The result is given in Figure 3. The dashed line is $y = x$.

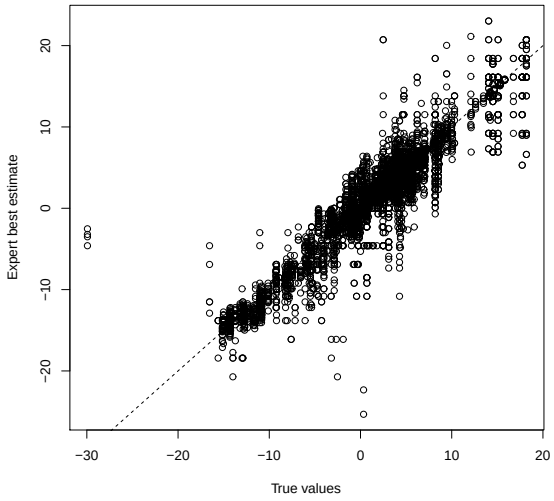

Figure 3: Scatter plot of the log of expert medians against the log of the true values for the seed variables from the Delft studies. The dashed line is $y = x$.

We see that there is certainly a relationship between the two quantities, with expert medians increasing as the true values of the seed variables increase. The expert assessment are also scattered both sides of the line $x = y$ indicating that experts both over- and under-estimate the true seed values. In fact, 46.2% of expert assessments are above the true value and 53.8% are below.

We can also view the distributions of the logarithms of the expert assessments and the true values of the seed variables. They are given as histograms in Figure 4.

We see that the distributions are taking broadly the same ranges and have similar shapes, although the expert medians appear to have a larger variance. The variance of the logarithm of the true seed variables is 30.79 and the variance of the logarithm of the expert medians is 32.82.

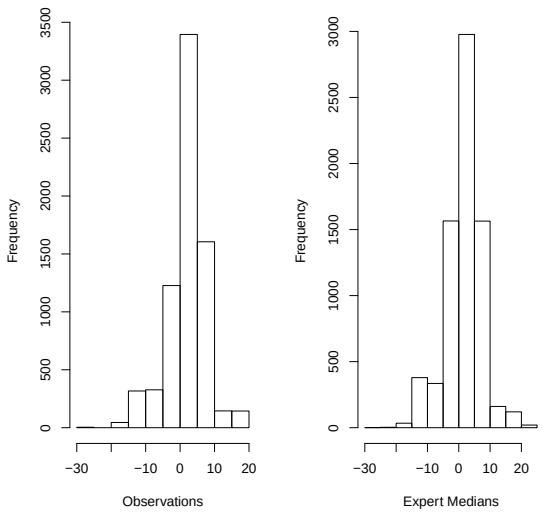

Figure 4: Histograms of the distributions of the log value of the seed variables and the log expert assessments.

We can also view the distributions of these quantities overlayed as density plots and this is given in Figure 5. The true values of the variables are given by the solid line and the expert medians are given by the dashed line.

We see that the shapes of the two distributions are very similar and it would appear to be not unreasonable to say that the expert medians follow the same distribution as the true values of the seed variables overall.

So far we have considered only the medians assessed by the experts for the seed variables. As part of the Delft database the experts gave their upper and lower 5% quantiles of their uncertainty for each of the seed variables. If the experts are assessing the uncertainty in the seed variables accurately, 5% of the seed variables should fall below the 5% quantile, 45% should fall between this and the median, 45% between the median and the 95% quantile and 5% above this. This is the basis of the calibration score in the Classical method [8].

We call these four intervals bins 1 to 4 respectively. A plot of the proportion of the experts' assessments falling into these bins (dark grey) against the theoretical proportions which should fall into each bin (light grey) is given in Figure 6.

We see that the proportions of true values falling into the central bins is too small whereas the proportion falling outside the 5% and 95% quantiles is too high. The actual proportions falling into the four bins are $(0.19, 0.27, 0.25, 0.29)$. This shows that overall the experts have been over-confident and have assessed 90% uncertainty bounds which are too narrow. This agrees with the findings

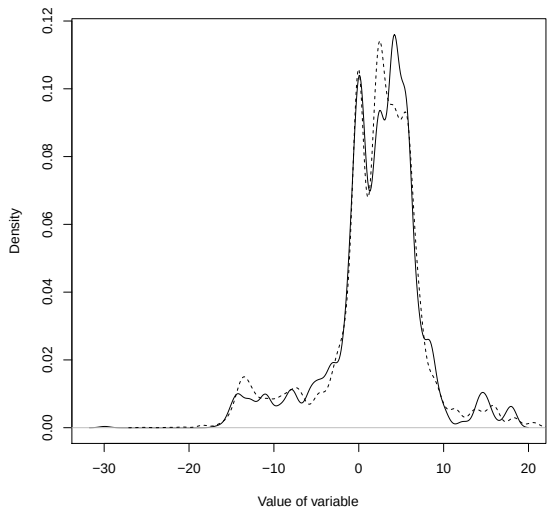

Figure 5: A density plot of the distributions of log value of the seed variables and log expert assessments.

of [16, 3, 17] and others.

5.2 Dependence in the studies

Initially we consider the correlations between experts. All correlations between experts are conditional on the true value of the seed variable and depend on the scale of the variable if different scales are used for seed questions. Suppose the median for expert i on question j of study k is $m_{i,j,k}$ and the true value of the seed variable is $v_{i,jk}$. Then we measure the correlations between the errors in the assessments of the experts, that is, conditional on the true value of the seed variable,

$$\epsilon_{i,j,k} = \frac{m_{i,j,k} - v_{i,j,k}}{v_{i,j,k}}.$$

The true values of the seed variables are often on very different scales, even within a single study, and even rescaling by the true value of the variable may not completely negate the effects of this scaling. Therefore, we will use both Pearson's coefficient and Kendall's Tau as our measures of correlation between different experts. Further justification in the case of Kendall's Tau is given in [20].

For each of the 45 studies we calculate the Pearson and Kendall correlations between each pair of experts. We then find minimum, lower quartile, median, upper quartile and maximum correlations for all experts for each individual

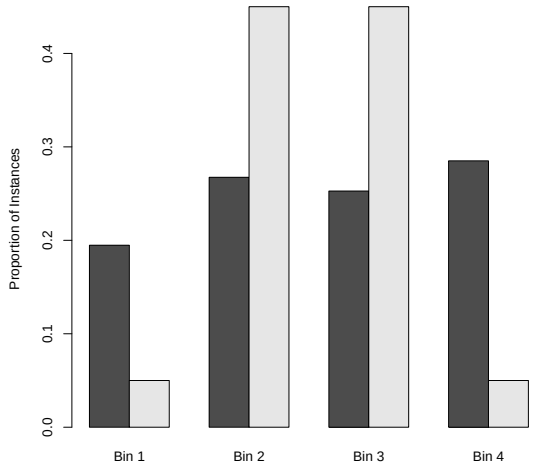

Figure 6: The observed proportion of seed variable falling into each of the bins (dark grey) and the theoretical proportions if experts were assessing the uncertainty correctly (light grey).

study and this is plotted for all studies as boxplots in the left and right hand sides respectively of Figure 7.

We see that there are a range of different correlations between experts in the studies using both Pearson and Kendall correlations. Most often correlations seem to be positive and some of the maxima are very close to 1. There are also some strong negative correlations. The range of correlations between experts are also very different for different studies, although studies used varying numbers of experts and seed questions. The correlations using Kendall's Tau seem more stable than those using Pearson correlation.

Of particular interest are strongly correlated experts. As strong is a subjective term we investigate correlations of above $(0.67, 0.75, 0.9, 0.95)$ and their negative counterparts. We are interested in the studies in which there is at least one combination of experts who are correlated to this degree as they are studies in which the usual assumption of independence between experts may not be suitable. Table 1 shows the proportions of studies in which this is the case for the various correlations.

We see from both the Pearson and Kendall correlations that there are many highly positively correlated experts across the studies. 93% and 73% of the studies have experts who are more than two thirds correlated using Pearson and Kendall correlations respectively. This is important as these experts are offering similar knowledge or are suffering from similar biases and so methods which assume independence between errors in the assessments of the experts

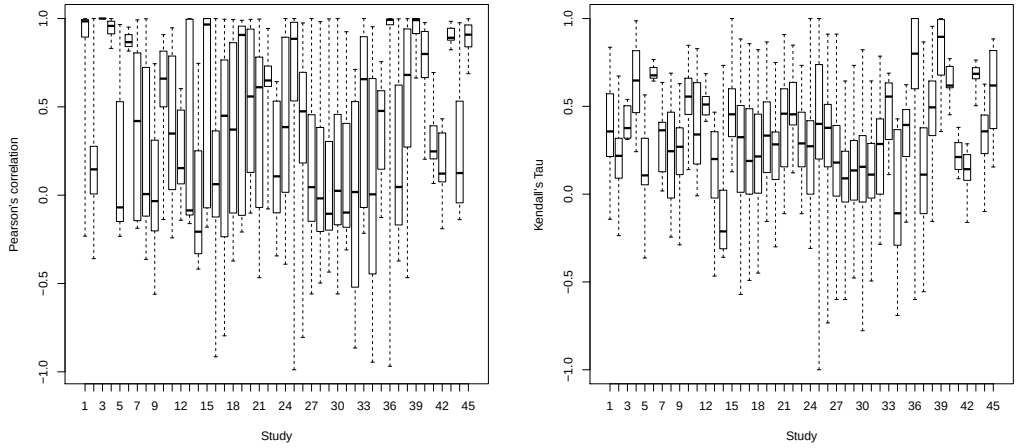

Figure 7: Boxplots representing the Pearson (left) and Kendall (right) correlations for each of the 45 Delft expert judgement studies.

Table 1: Table showing the proportion of studies with at least one pair of experts who are highly correlated.

	Pearson		Kendall	
	Positive	Negative	Positive	Negative
0.67	0.93	0.16	0.73	0.09
0.75	0.84	0.16	0.56	0.04
0.90	0.82	0.09	0.22	0.02
0.95	0.73	0.04	0.16	0.02

may result in a forecast which is over-confident. That is, they over-estimate the amount of unique information we recieve from the experts and so our uncertainty on the unknown is reduced by too much. Again there are differences between the two measures with the Pearson method displaying more very high correlations than the Kendall method.

There are far fewer studies which contain highly negatively correlated experts, although they do exist for all correlations in Table 1. It is important to incorporate negative dependence of experts into aggregation methods as negatively correlated experts may be giving complementary information and, if independence is assumed, may result in an answer which is under-confident. That is, they under-estimate the amount of unique information we receive from the experts and so our uncertainty on the unknown is not reduced enough.

We can also consider the correlations through time in individual experts within studies. To do so, we calculate the correlations for different questions within each expert in each study. We again calculate Pearson correlation conditional on the true value of the variable. We can plot the distribution of the correla-

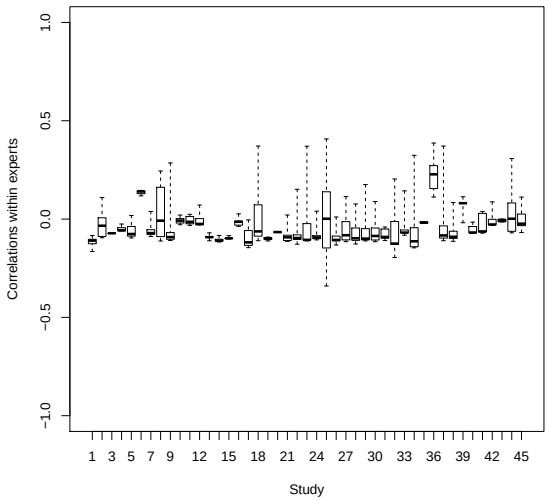

Figure 8: Boxplots of the distributions of correlations within experts in each of the 45 TU Delft studies.

tions for experts in each study as a boxplot and this is given for all studies in Figure 8.

We see that the boxplots are much more closely centred towards zero indicating that within expert correlations are not as prevalent as between expert correlations in the studies. There also appear to be very few experts displaying large positive and negative correlations between questions. This could be a cause for concern for those methods which utilise calibration questions to weight experts, as it suggests that an expert who is able to predict a single unknown well, is not necessarily going to outperform other experts for a different unknown.

As a result of the lack of strong correlations within experts for different questions, correlations between experts for different questions will not be considered.

6 Comparison of Aggregation Approaches

6.1 Methods of Comparison

In order to compare the different mathematical aggregation techniques considered we will need to define suitable comparison measures. As many of the studies in the Delft database have a small number of seed questions we will use all seed variables to calculate the expert weights for each technique. These

variables will also be used to calculate the dependence between experts for the techniques which include these measures. The mathematical aggregation techniques will then be assessed on the ability of the aggregated judgements to estimate the true values of the seed variables. That is, in this paper we consider in sample validation. To ensure that the results are robust, for each analysis we shall also perform cross validation via leave one out analysis. That is, we will fit the model to all but one of the seed variables and use this to predict the remaining seed variable. We then repeat this for all of the seed variables in a study. The cross validation results will be given in brackets following the in sample validation results in text and tables.

The between expert correlations will be modelled using the multivariate Normal and copula approaches as it was these that were found to be prevalent in the expert judgement studies.

To assess the accuracy of the point estimates from the aggregation methods we shall use three criteria based on the proportion errors in the estimates $(\hat{y}_{i,j} - y_i)/y_i$. The three model fit criteria we use to assess the models are the mean absolute proportion error (MAPE), $1/N \sum_{i=1}^{N} |(\hat{y}_{i,j} - y_i)/y_i|$, the root mean squared proportion error (RMSPE) $\{1/N \sum_{i=1}^{N} [(\hat{y}_{i,j} - y_i)/y_i]^2\}^{0.5}$, and the maximum absolute proportion error (MAXPE), $\max_{i \in 1,\ldots,N} |(\hat{y}_{i,j} - y_i)/y_i|$. We will use these metrics to compare the techniques for each study and across studies.

6.2 Comment on Modelling

In order to perform the analysis, decisions needed to be made with regards to both the data and the individual models. We will set these out in this section. In order to enable comparison of all studies, decisions were taken for all studies and hence, for a careful analysis in an individual study, each method could potentially provide a better fit. However, the current analysis compares the performance of an "off the shelf" version of each model. Initially we consider the data. There were cases of experts not giving values for all of the seed variables. These experts were removed from the analysis.

For the Multivariate Normal model, a flat prior was used for the decision maker. The correlations between the experts were empirically estimated using the data from the seed variables from the study in question. For the copula model, experts were assumed exchangeable and the correlation between experts was assessed from the data. The Joe copula was used between pairs of experts. Log-normal prior distributions were used for individual variables.

For the Classical model, the optimal decision maker using both global and item weights was calculated and the better of these was taken for each perfor-

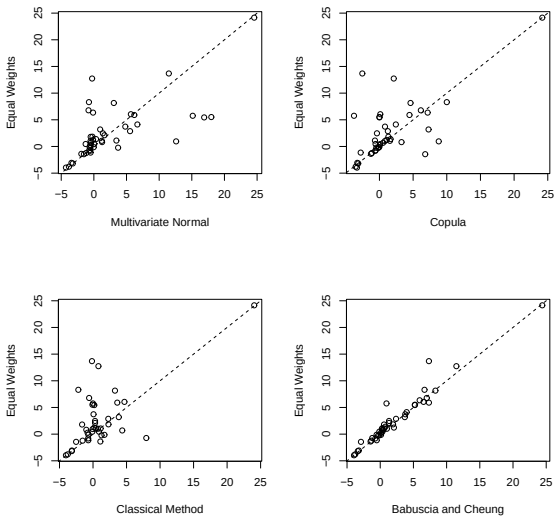

Figure 9: Scatter plots comparing the MAPE from Multivariate Normal, Copula, Classical and Babuscia and Change methods to equal weights.

mance measure in each study. For the Babuscia and Cheung method only the calibration score was calculated and the weights for experts were made based on this. That is, the probabilistic thinking score could not be calculated in this case as the data are historical.

6.3 Results

Initially we use each of the methods to calculate the aggregated prediction, variously the mean or median, for each seed question in each study in the investigation. We use equal weighting of the experts as the baseline to compare the models against. The first question is therefore; are each of the methods outperforming equal weighting of the experts for the studies?

In Figure 9 we give scatter plots of the MAPE for each of the four aggregation methods against equal weighting. Any points above the dashed $y = x$ line indicate that the aggregation approach is superior for that study.

We see that, though there are points below and above the line in all plots, each method is outperforming equal weight more than half of the time. The exact proportions for the Multivariate Normal, Copula, Classical and Babuscia and Change methods are 0.60 (0.56), 0.53 (0.56), 0.62 (0.55) and 0.67 (0,58) respectively. The equivalent scatter plots for RMSPE and MAXPE show greater superiority over equal weights. The proportions of studies where each of the methods are outperforming equal weights are, in the order above, 0.67 (0.8), 0.60 (0.64), 0.64 (0.6) and 0.69 (0.58) for RMSE and 0.69 (0.67),0.69

	MVN	Copula	Classical	Babuscia	Equal
MAPE	0.38 (0.44)	0.22 (0.18)	0.27 (0.16)	0.09 (0.18)	0.04 (0.04)
RMSPE	0.40 (0.60)	0.24 (0.18)	0.24 (0.13)	0.09 (0.09)	0.02 (0.00)
MAXPE	0.42 (0.38)	0.33 (0.29)	0.18 (0.20)	0.07 (0.13)	0.00 (0.00)

Table 2: The proportion of studies in which each method performs best for each of the criteria

(0.61),0.67 (0.68), and 0.69 (0.60) for MAXPE. In general, there isn't much to choose from between the four methods on this metric. The cross validation results are consistent with the in sample results.

We can also identify the best performing method under our three criteria for each of the studies. The results of this are give in Table 2.

We see that, for all three criteria, the Multivariate Normal (MVN) method is providing the best predictions the largest proportion of the time. Both the Copula and Classical approaches also have good performance over the studies. While the Babuscia method typically outperforms equal weights, it is rarely the optimal method under any of the three criteria. Equal weighting is very rarely the optimal weighting strategy. The cross validation results show a slightly greater superiority for the Bayesian methods.

We can also evaluate the differences in the performance of the models in studies with high dependence between experts. We consider high dependence studies to be those in which at least one pair of experts have a Kendall's Tau of at least 0.75. There are 25 such studies. The scatterplots for each of the methods showing their performance on MAPE compared to equal weights for each of these studies are given in Figure 10.

In this case we see that the copula method appears to be at least as good as equal weights on almost all occasions. The equivalent scatter plots for RMSPE and MAXPE show greater superiority over equal weights. The proportions of the high dependency studies in which the methods are outperforming equal weights on MAPE are 0.60 (0.60), 0.48 (0.48), 0.52 (0.44) and 0.76 (0.48) for the Multivariate Normal, Copula, Classical and Babuscia and Change approaches respectively. For RMSPE these proportions are 0.72 (0.76), 0.56 (0.60), 0.56 (0.44) and 0.76 (0.48) and for MAXPE they are 0.76 (0.72), 0.68 (0.72), 0.60 (0.56) and 0.76 (0.52). Overall it would appear that all methods are still outperforming equal weights on average. Interestingly here, the cross validation shows a greater discrepancy between the Bayesian and opinion pooling methods than the in sample validation.

We can also calculate the best performing method for each of these studies under each of the three criteria. We give the results in Table 3.

From the table we see the dominance in these studies of the methods which

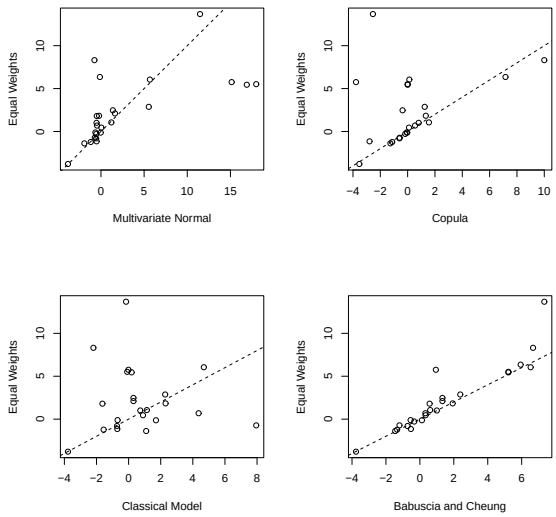

Figure 10: Scatter plots comparing the MAPE from Multivariate Normal, Copula, Classical and Babuscia and Cheung methods to equal weights for high dependency studies.

	MVN	Copula	Classical	Babuscia	Equal
MAPE	0.32 (0.40)	0.28 (0.24)	0.24 (0.16)	0.08 (0.12)	0.08 (0.08)
RMSPE	0.36 (0.6)	0.32 (0.24)	0.2 (0.12)	0.08 (0.04)	0.04 (0.00)
MAXPE	0.44 (0.36)	0.44 (0.44)	0.04 (0.16)	0.08 (0.04)	0.00 (0.00)

Table 3: The percentage of high dependency studies that each method performs best for each of the criteria

include correlations between the experts. That is, in all studies the methods including dependency between experts performed best 60% of the time for MAPE, 64% of the time for RMSPE and 75% of the time for MAXPE but in high dependency studies these numbers rise to 60% of the time for MAPE, 68% of the time for RMSPE and 88% of the time for MAXPE. It would appear that incoprorating dependency can lead to improved model performance. In the case of MAPE, though the models including dependency are optimal the same percentage of the time, the methods assuming independence are optimal 32% of the time rather than 36% of the time for all studies. These trends can also be seen in the cross validation results.

6.4 Main Findings and implications for forecasting

The analyses conducted in Sections 5 and 6 have produced notable findings which deserve some elaboration in the context of current debates within the expert judgement and forecasting literatures. In particular, we consider them with respect to Bolger and Rowe's comparisons of mathematical aggregation (MA) and behavioural aggregation (BA) [5] and accompanying discussants and response [9, 35, 33, 6]. We first emphasise that we cannot compare MA to BA as we only treat methods for MA in this paper. We are also unable to contribute to the debate put forward by [33] as to whether any aggregation of judgements is desirable.

In Section 5 we observed overconfidence in the overall proportions of uncertainty assessments from experts falling within their 90% uncertainty limits. This is consistent with the views put forward by Winlker [35] that "overconfidence is an important issue in subjective probability forecasts, leading to ... probability distributions that are too tight" and Bolger and Rowe [6] that "many experts are overconfident". Cooke [9] points out that not all experts are overconfident and this is certainly also true in the studies analysed.

We observed very strong positive correlations for errors in point predictions between experts in expert judgement studies, and very few strong negative correlations. We found there to be comparatively few correlations, both positive and negative, in the errors in prediction for individual experts over multiple predictions. These results have various implications as weights in MA techniques are often sensitive to high dependence among the forecast errors [35]. The lack off correlations within experts could be related to the assertion [6] that accuracy in expert predictions "is not stable over experts and situations". More comprehensive work is needed in this area to investigate this further.

In Section 6, our comparison of mathematical aggregation methods indicated that all methods out-performed equal weighting of experts a majority of the

time when considering point predictions. The methods which included correlations between the errors in prediction from experts performed better overall than those which didn't, and this preponderance showed a modest increase as we included only the studies in which there were highly correlated experts. This would appear to be in conflict with the view of [5] that "no significant benefits are likely to accrue from unequal weighting in mathematical weighting". It would appear that, in the studies considered here, the MA techniques are able to increase the accuracy of predictions over equal weighting, at the expense of additional effort.

We have seen that dependency between experts can affect the accuracy of estimates resulting from aggregation models. In particular, if there is dependency present, there is some evidence that assessments of unknowns are improved by choosing a model which takes these into account. This is in contrast to findings in forecasting from, for example, [27, 32]. Similarly, in our analysis we found that more complex aggregation models outperformed simple averaging. Again, this is in contrast to a body of previous work, including that of [18, 7].

The specific judgemental techniques considered in this paper could have particular utility in forecasting. A much considered problem in forecasting is the combination of multiple data driven forecasts. De Menezes *et al* [12] review the techniques used to do so and conclude that it is important to consider the uncertainty, rather than simply the point estimate, when deciding on the method to combine. We would argue from the results in this paper that, if we wish to combine forecasts from experts for a quantity on which we have no or little relevant data, then a combination of the assessments in Section 6.1 and and a measure of uncertainty should be used to choose the aggregation technique. Of course, this assumes that we wish to use mathematical rather than behavioural aggregation, which may not always be the case.

One advantage forecasting has over other areas in which expert judgement is used, is that often the true values of the unknowns will be revealed at some point in the future. For example, the near future local climate behaviour in [14] and the unemployment and inflation rates of [19]. The result is that the choice of seed questions, one of the most difficult tasks in mathematical aggregation, could reduce to asking questions about values to be revealed sooner than those of interest. This could potentially improve the performance of mathematical aggregation approaches in comparison to studies in, for example, risk analysis, where probabilities are typically elicited but only one reality is observed.

In financial forecasting, most interest lies in the extremes of the data, in the peaks and troughs of financial performance. Unfortunately, this is where data is most scarce and forecasts from widely used models such as the Gaussian copula have been shown to be very poor [30] as a result of the presence of

tail dependence. Expert judgement aggregation could play an important role in improving forecasts under multiple types of dependence in such cases, as the judgements of a single expert are unlikely to be judged sufficient in such cases..

7 Summary and Discussion

Overall, we conclude that there is evidence of strong dependencies between experts in expert judgement studies but little evidence of within expert correlations. Further, we found some evidence that taking these correlations into account in models could improve forecasts.

While there are models in the Bayesian literature which take both of these types of dependence into account, there is to date no method which includes all of the correlations identified in Section 4.1.

This paper has considered dependency within and between experts in expert judgement studies. Another issue of importance is that of eliciting dependencies between unknowns from experts. This is of particular relevance in forecasting, where correlations at both a single point in time and through time could have strong effects on the accuracy of forecasting models.

References

[1] J. Armstrong. Principles of forecasting: Combining forecasts. *International Series in Operations Research & Management Science*, 30:417–439, 2001.

[2] A. Babuscia and K. Cheung. An approach to perform expert elicitation for engineering design risk analysis: methodology and experimental results. *J. Royal. Stat. Soc.: Series A*, 177:475–497, 2014.

[3] S. Barclay and C. Peterson. Two methods for assigning probability distributions. Technical Report Dt/TR 75-4, Decisions and Designs, Inc.

[4] P. Barndt, J. Freeman, and P. Schrodt. Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30:944–962, 2014.

[5] F. Bolger and G. Rowe. The aggregation of expert judgement: do good things come to those who weight? *Risk Analysis*, 35:5–26, 2015.

[6] F. Bolger and G. Rowe. There is data, and then there is data: Only experimental evidence will determine the utility of differential weighting of expert judgment. *Risk Analysis*, 35:21–26, 2015.

[7] R. Clemen. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5:559–583, 1989.

[8] R. Cooke. *Experts in Uncertainty*. Oxford: Oxford University Press, 1991.

[9] R. Cooke. The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 35:12–15, 2015.

[10] R. Cooke and L. Goossens. TU Delft expert judgement database. *Reliability Engineering and System Safety*, 93:657–674, 2007.

[11] R. Cooke and E. Jager. A probabilistic model for the failure frequency of gas pipelines. *Risk Analysis*, 18:511–527, 1998.

[12] L. de Menezes, D. Bunn, and J. Taylor. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120:190–204, 2000.

[13] M. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974.

[14] A. DeWispelare, L. Herren, and R. Clemen. The use of probability elicitation in the high-level nuclear waste regulation program. *International Journal of Forecasting*, 11:5–24, 1995.

[15] S. French. Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas*, 105:181–206, 2011.

[16] P. Garthwaite, J. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100:680–700, 2005.

[17] P. Garthwaite and A. O'Hagan. Quantifying expert opinion in the UK water industry: An experimental study. *The Statistician*, 49:455–477, 2000.

[18] D. Hendry and M. Clements. Pooling of forecasts. *Econometrics Journal*, 7:1–31, 2004.

[19] M. Jochmann, G. Koop, and R. Strachan. Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks. *International Journal of Forecasting*, 26:326–347, 2010.

[20] M. Jouini and R. Clemen. Copula models for aggregating expert opinions. *Operations Research*, 44:444–457, 1996.

[21] D. Kahneman and A. Tversky. Subjective probability: A judgement of repetitiveness. *Cognitive Psychology*, 3:430–454, 1971.

[22] N. Kerr and R. Tindale. Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting*, 27:14–40, 2011.

[23] D. Lindley, A. Tversky, and R. Brown. On the reconciliation of probability judgements (with discussion). *J. R. Stat. Soc. A*, 142:146–180, 1979.

[24] S. Makridakis and M. Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16:451–476, 2000.

[25] C. Min and A. Zellner. Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics*, 56:89–118, 1993.

[26] P. Morris. Combining expert judgements: A Bayesian approach. *Management Science*, 23:679–693, 1977.

[27] P. Newbold and C. Grainger. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A*, 137:131–149, 1974.

[28] A. O'Hagan, C. Buck, A. Daneshkhah, J. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, 2006.

[29] G. Rowe and G. Wright. The delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, 51:353–375, 1999.

[30] S. Shreve. Did mathematical models cause the financial fiasco? *Analytics magazine*, pages 6–7, 2009.

[31] P. Slovic. From Shakespeare to Simon: Speculation - and some evidence - about man's ability to process information. *Oregon Research Bulletin*, 12, 1972.

[32] A. Timmermann. Forecast combinations. *Handbook of Economic Foracasting*, 1:135–196, 2006.

[33] M. Winkler. Our knowledge of the world is often not simple: Policymakers should not duck that fact, but should deal with it. *Risk Analysis*, 35:19–20, 2015.

[34] R. Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27:479–488, 1981.

[35] R. Winkler. Equal versus differential weighting in combining forecasts. *Risk Analysis*, 35:16–18, 2015.